

TextGrid-Kerncodierung

AG Textformate

15. Juni 2007

1 Wozu Kerncodierungen?

Texte von Projekten, die in TextGrid veröffentlicht werden, sollen auf zwei verschiedene Weisen durchsuchbar sein:

1. eine *projektspezifische Suche* durchsucht nur Texte jeweils eines Projekts, die Projektmacher können jedoch Suchmasken und Ergebnisdarstellung auf die ganz spezifischen Besonderheiten ihres Projekts anpassen;
2. die *TextGrid-übergreifende Suche* bietet eine Suchmöglichkeit über alle in TextGrid veröffentlichten Texte hinweg.

Für ersteres können die Projekte alle Freiheiten der TEI-Codierung (oder sogar darüberhinausgehende XML-Codierungen) nutzen. Für letzteres ist es jedoch nötig, die projektspezifischen Codierungen auf eine gemeinsame *Kerncodierung* abzubilden. Diese Abbildung wird durch einen *Adaptor* (etwa ein XSLT-Stylesheet) realisiert.

2 Anforderungen an die Kerncodierungen

Wir haben bislang drei Anforderungen an diese Kerncodierung ausgemacht:

1. **Intelligente Suche.** Anders als die bloße Freitextsuche soll die Suche mittels textsortenspezifischen Kodierungen genauere und damit intelligentere Abfragen ermöglichen. Grundlage ist hierfür die Frage, welche Aspekte der Textsorte in einer textsortenübergreifenden Suche und in einer textsortenspezifischen, aber einzeltext- und projektübergreifenden Suche von besonderem Interesse sind (wir zielen darauf ab, den größeren Teil der Suchen zu unterstützen, aber nicht alle und nicht sehr spezielle). Ein besonderes Problem stellen hierbei die linguistischen Korpora, da die Benutzerbedürfnisse hier sehr spezifisch sind.
2. **Repräsentation der Suchergebnisse.** Hier gibt es zwei Anforderungen:
 - a) die Position des Suchergebnisses im editorischen Kontext (Goethe→Werther→Brief vom 13. May)
 - b) die notwendigen typographischen Konventionen der Textsorten (z.B. Verse als einzelne Zeilen, typografische Differenz zwischen Regieanweisung und Figurentext)
3. **Automatische Weiterverwendung** und -verarbeitung von Texten, z. B. zur Extraktion von Zusammenhängen in Wörterbuchnetzen.

Wir haben den Ansatz gewählt, diese Kodierungen nach Funktion zu modularisieren: Die Kerncodierung umfasst

1. Metadaten,
2. allgemeine Strukturdaten
3. allgemeine übergreifende Daten (etwa Abkürzungen)
4. Textsortenspezifische Kodierungen:
 - Wörterbücher
 - Linguistische Korpora
 - Prosa
 - Dramen
 - Lyrik
 - Literaturwissenschaftliche Editionen
 - Briefe

Es ist noch offen, welche weiteren Textsorten wir in TextGrid 1 unterstützen wollen.

Wenn die getrennt entwickelten Module vorliegen, werden wir sie integrieren, um auch gemischte Textsorten zu unterstützen.

3 Technische Bestandteile einer Kerncodierung

Zur Spezifikation einer Kerncodierung gehören:

1. ein *formales Schema* auf der Basis von TEI und ggf. verwandten Standards, das die Kerncodierung spezifiziert und in ODD, Relax NG oder W₃C XML Schema ausgeführt ist
2. *Dokumentation* zu diesem Schema, die insbesondere erläutert
 - welche Suchanfragen bzw. Darstellungsaufgaben durch die einzelnen Schemaelemente befriedigt werden sollen
 - wie das Schema zu verwenden ist
3. Beispiele (vorher/nachher) von projektspezifisch und kerncodierten Texten der entsprechenden Textkategorie
4. ein Beispieladaptor.

4 Was die Kerncodierung nicht ist

Die Kerncodierung stellt in mancher Hinsicht einen Kompromiss dar, um auch vorhandene Projekte möglichst nicht auszuschließen. Deshalb ist die Kerncodierung explizit *nicht* als Best-Practice-Empfehlung für die Codierung neuer Projekte zu verstehen, sondern allein zur Realisierung projektübergreifender Funktionen in TextGrid gedacht. Auch Projekte, die TextGrid nutzen wollen, sollten für ihre Texte weiterhin eine Codierung wählen, die ihren spezifischen Anforderungen entspricht.